Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

ECOVISOR: AN AI-DRIVEN FRAMEWORK FOR SUSTAINABLE RESOURCE ALLOCATION IN CLOUD DATA CENTERS

Smita Ketan Hadawale and Snehal Dinesh Dubey

Department of Computer Science,

Pillai College of Arts, Commerce and Science (Empowered Autonomous), New Panvel

**Corresponding author E-mail: smitahadawale@mes.ac.in*

Received: 10 January 2026	Revised: 15 February 2026	Accepted: 17 March 2026	Published: 16 April 2026

DOI: **https://doi.org/10.5281/zenodo.19609752**

Abstract:

Global data centers consume approximately 200–250 terawatt-hours (TWh) of electricity annually, representing about 1–1.5% of global energy demand and 0.3% of worldwide carbon dioxide (CO_2) emissions. With artificial intelligence (AI) workloads expected to double energy consumption by 2026, improving the sustainability of cloud infrastructure has become critical. Existing resource allocation mechanisms remain inefficient, with average server utilization around 25–30% and significant over-provisioning. This paper proposes EcoVisor, an AI-driven framework for sustainable resource allocation in virtualized cloud environments. EcoVisor combines a hybrid workload forecasting model integrating LightGBM and GRU networks with a multi-objective reinforcement learning agent based on Proximal Policy Optimization (PPO). The framework dynamically optimizes CPU, memory, and virtual machine placement while minimizing energy consumption, carbon emissions, water usage, and hardware stress. Simulation and testbed evaluations indicate potential reductions of 20–30% in energy consumption and significant environmental impact without performance degradation.

Keywords: *Data Centers, Energy Efficiency, Cloud Computing, Artificial Intelligence, Resource Allocation, Sustainability,*

1. Introduction

Data centers are the backbone of modern digital infrastructure but pose significant environmental challenges. According to the International Energy Agency (IEA), they consume approximately 200–250 TWh of electricity annually, accounting for about 1–1.5% of global electricity demand [1], and contribute nearly 0.3% of worldwide CO_2 emissions, comparable to the aviation industry [2]. The rapid expansion of artificial intelligence (AI) applications has intensified these concerns, as training large-scale models requires substantial computational

resources, increasing energy use and emissions. Studies suggest that data center energy demand may double by 2026 due to AI-driven workloads [3].

Despite advances in virtualization, resource utilization remains inefficient. Average server utilization is typically only 25–30%, while idle servers may consume up to 60% of peak power due to over-provisioning strategies aimed at maintaining reliability [4]. Artificial intelligence offers a solution through predictive and adaptive resource management. This study proposes EcoVisor, an AI-driven framework combining hybrid workload forecasting, reinforcement learning, and environmental-aware scheduling to enhance sustainability while preserving performance.

2. Related work

Previous studies have investigated energy-efficient strategies for cloud data centers, including dynamic voltage and frequency scaling (DVFS), workload consolidation, and thermal-aware scheduling [5]. Cloud orchestration systems such as Kubernetes and OpenStack have also introduced auto-scaling mechanisms to improve resource utilization.

Recent research has explored machine learning approaches for workload prediction. Gradient boosting algorithms such as LightGBM have demonstrated strong performance for structured data prediction tasks due to their efficiency and scalability [6]. Recurrent neural networks, particularly GRU models, have been widely used for time-series forecasting because they capture temporal dependencies in sequential workloads [7].

Reinforcement learning has also been applied to cloud resource management. Algorithms such as Proximal Policy Optimization (PPO) provide stable policy learning for complex decision-making environments [8]. However, most existing frameworks optimize only single objectives such as energy consumption or performance.

EcoVisor advances previous work by combining hybrid workload forecasting with multi-objective reinforcement learning while integrating environmental factors such as carbon intensity and cooling efficiency.

3. Methodology

3.1 research design

The study adopts the Design Science Research (DSR) methodology, which is widely used in information systems research to develop and evaluate innovative technological artifacts [9]. The research process includes the following phases:

- Problem identification
- Solution design
- Artifact development
- Experimental evaluation
- Research communication

3.2 Experimental setup

The EcoVisor framework was evaluated across multiple environments to ensure robustness and scalability.

i. Simulation Environment: CloudSim was used to simulate up to 10,000 servers, enabling large-scale evaluation of resource allocation algorithms [10].

ii. Physical Testbed: A cluster of 20–30 servers was used to measure real-world system overhead and operational performance.

iii. Production-Like Environment: Integration testing was conducted using OpenStack and Kubernetes with 100–200 virtual machines.

iv. Industry-Scale Validation: Real-world experiments were conducted on a 40-server rack environment to validate system behavior under realistic workloads.

3.3 Datasets

Workload datasets were obtained from publicly available cloud traces widely used in cloud research.

- Google Cluster Trace 2019: Contains workload traces from over 12,500 machines across 29 days [11].
- Alibaba Cluster Trace 2018: Includes operational workload data from 1,300 servers over 8 days [12].
- Azure Public Dataset 2022: Contains traces of approximately 2 million virtual machines operating over 30 days [13].

Environmental data sources include:

- Electricity grid carbon intensity datasets
- Meteorological datasets from NOAA
- Server power consumption models from SPECpower
- Water stress indicators from WRI Aqueduct

3.4 Ecovisor Framework

The EcoVisor architecture consists of three primary components.

A. Hybrid workload forecasting

The forecasting model combines:

- LightGBM for feature-based workload prediction
- GRU networks for sequential workload pattern analysis

Hybrid models have been shown to improve prediction accuracy by combining statistical learning with temporal pattern recognition [6][7].

B. Reinforcement learning optimization

A multi-objective reinforcement learning agent based on PPO dynamically allocates cloud resources.

The state space includes:

- CPU utilization
- Memory usage
- Pending workloads
- Carbon intensity levels

The reward function balances:

- System performance
- Energy consumption
- Carbon emissions
- Hardware utilization

PPO was selected because it provides stable training performance and has been successfully applied in large-scale decision systems [8].

C. Carbon-aware and thermal-aware scheduling

EcoVisor schedules workloads based on:

- Renewable energy availability
- Real-time grid carbon intensity
- Data center thermal distribution

This approach helps reduce both cooling energy demand and carbon emissions, aligning data center operations with renewable energy availability.

4. Evaluation metrics

The EcoVisor framework was evaluated using operational, environmental, and system-level metrics.

A. Operational metrics

- Response time
- Throughput
- SLA violation rate

B. Environmental metrics

- Energy consumption reduction
- Carbon emission reduction
- Water usage effectiveness (WUE)
- Server utilization rate

C. System metrics

- Forecasting accuracy (MAPE)
- Decision latency
- CPU and memory overhead

5. Results

Experimental evaluation demonstrated significant improvements in resource utilization and sustainability metrics.

Key results include:

- Energy reduction: 20–30%
- Carbon emission reduction: 25–35%
- Water usage reduction: 20–30%
- Server utilization improvement: from 25% to 45–55%
- Hardware reduction: approximately 20% fewer servers required

Decision latency remained below 100 ms, ensuring real-time resource allocation without affecting application performance.

6. Discussion

The results indicate that AI-based resource management can significantly improve the sustainability of data centers. Unlike traditional reactive scaling approaches, EcoVisor employs predictive optimization, allowing the system to allocate resources proactively based on workload forecasts.

The integration of environmental metrics into the reinforcement learning reward function enables simultaneous optimization of energy consumption, carbon emissions, and hardware utilization. Furthermore, explainable AI techniques enhance transparency, enabling administrators to understand and trust automated decisions.

7. Limitations

Several limitations should be acknowledged:

- Models trained on hyperscale datasets may require adaptation for smaller enterprise environments
- Carbon-aware scheduling depends on availability of real-time grid carbon intensity data
- Water savings depend on the cooling infrastructure used in data centers

- Global adoption projections assume favorable economic and regulatory conditions

8. Future work

Future research directions include:

- Transfer learning to adapt models across different data center environments
- Integration with renewable energy storage systems
- Expansion to edge and distributed computing environments
- Predictive hardware maintenance using machine learning
- Federated learning approaches across multiple data centers

Conclusion

This study presented EcoVisor, an AI-driven framework for sustainable resource allocation in cloud data centers. By integrating hybrid workload forecasting, multi-objective reinforcement learning, and carbon-aware scheduling, EcoVisor significantly improves resource utilization while reducing environmental impact.

The results demonstrate that AI can transform data centers into more sustainable computing infrastructures, supporting global efforts toward environmentally responsible digital systems.

References

1. International Energy Agency. (2023). *Data centres and data transmission networks*.
2. Masanet, E., *et al.* (2020). Recalibrating global data center energy-use estimates. *Science*.
3. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
4. Shehabi, A., *et al.* (2016). *United States data center energy usage report*. Lawrence Berkeley National Laboratory.
5. Beloglazov, A., & Buyya, R. (2012). Energy-efficient resource management in virtualized cloud data centers. *Future Generation Computer Systems*.
6. Ke, G., *et al.* (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*.
7. Cho, K., *et al.* (2014). Learning phrase representations using RNN encoder–decoder. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
8. Schulman, J., *et al.* (2017). *Proximal policy optimization algorithms*. arXiv.
9. Hevner, A., *et al.* (2004). Design science in information systems research. *MIS Quarterly*.
10. Calheiros, R., *et al.* (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments. *Software: Practice and Experience*.
11. Reiss, C., *et al.* (2019). *Google cluster-usage traces*. Google Research.
12. Alibaba Group. (2018). *Alibaba cluster trace program*.
13. Microsoft. (2022). *Microsoft Azure public dataset*.